Una propuesta para clasificación de roles de usuarios en foros de discusión técnicos

Nadina Martínez Carod, Gabriela Aranda, Valeria Zoratto, Christian Murray

Grupo GIISCo, Facultad de Informática, Universidad Nacional del Comahue
Buenos Aires 1400 (8300) Neuquén, Argentina
{nadina.martinez|gabriela.aranda|vzoratto}@fi.uncoma.edu.ar,
cristianmurray@hotmail.com

Resumen Este trabajo se enfoca en la detección de roles de usuarios participantes en foros de discusión técnicos, con el propósito de incorporar dicha clasificación como parte de una herramienta de recuperación de información de foros de discusión. El objetivo de dicha herramienta es, a partir de una consulta específica, determinar un ranking de hilos de discusión relevantes, seleccionados de varios foros de discusión. El ranking resultante estará ordenado según diferentes características de calidad, dentro de las cuales serán considerados los roles de los usuarios participantes. En este trabajo se definen las características de los distintos roles, la forma de clasificar los mismos y se presenta una validación preliminar del modelo propuesto en un caso de estudio.

1. Introducción

Los foros de discusión técnicos facilitan la creación de espacios de debate científico y permiten compartir conocimiento específico en ambientes informales, lo que los transforma en bitácoras actualizadas sobre temas específicos, que muy a menudo son utilizados para la recuperación de información (RI). Los usuarios que participan en dichos foros comparten conocimientos y experiencias, siendo éstas de gran utilidad cuando un problema surge repetidamente , por lo que el desafío fundamental radica entonces, en detectar los hilos de discusión que contienen las soluciones más adecuadas para un problema particular, analizando los debates que suceden en uno o varios foros de discusión a la vez. Con ese objetivo en mente, en [1] se definió un conjunto inicial de características de calidad para evaluar las soluciones encontradas en diferentes foros de discusión, seguido del diseño de un modelo para obtener un ranking de soluciones posibles, considerando diferentes ópticas. Por un lado se propuso un conjunto preliminar de métricas de calidad [2], el cual fue probado en un grupo de hilos de discusión reales [3]. Por otro lado, se planteó un modelo de clasificación basado en el análisis del texto de los hilos bajo estudio [4]. Posteriormente, se extendió dicho modelo mediante la incorporación de sinónimos según la estructura semántica de las oraciones [5], para lo cual se utilizó la base de datos léxica WordNet [6] de manera conjunta con la herramienta de procesamiento de lenguaje natural Stanford POS Tagger [7].

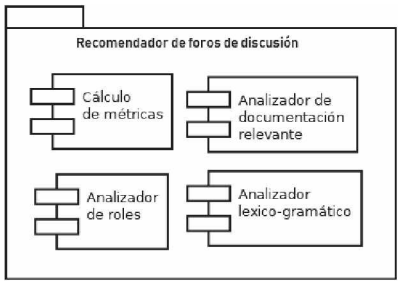

Figura 1. Esquema del recomendador de foros

Para extender la propuesta inicial de un recomendador de foros se ha agregado una nueva componente que utiliza el rol de los usuarios que participan compartiendo opiniones y experiencias en el foro, bajo la premisa que el rol del usuario está en estrecha relación con el conocimiento del mismo. Para entender la funcionalidad de la herramienta en construcción, en la Figura 1 se muestra el esquema del recomendador de foros de discusión, con las componentes mencionadas. A continuación se describirá más en detalle el agregado del componente *Analizador de roles*.

El resto del artículo está organizado de la siguiente manera: en la Sección 2 se introduce una herramienta que mantiene y gestiona la información contenida en foros de discusión para establecer el ranking de la información existente en dichos foros. Esto se realiza incorporando la funcionalidad de clasificar los participantes de los foros técnicos, para luego poder asociarles un grado o nivel de conocimiento en el tema. En la Sección 3 se valida la estrategia propuesta mediante la utilización de un caso de estudio. Finalmente, se analizan los resultados obtenidos comparándolos con una estimación de la experiencia de los participantes mediante una encuesta personal, se presentan las conclusiones y líneas de trabajo futuro.

2. Recomendador de foros

En [1] se presentó el modelo conceptual de la información contenida en foros de discusión desde el punto de vista del usuario externo, identificándose las entidades más importantes y sus atributos. Siguiendo este modelo, se plantea la arquitectura de la herramienta *Recomendador de foros* [8] cuyo fin es recomendar una lista de hilos de discusión técnicos, ordenada de acuerdo a determinada prioridad. Esta herramienta está diseñada en tres capas, como se muestra en la Figura 2.

En la capa de Datos se destaca el *Repositorio de hilos*, donde se mantienen los archivos recuperados en formato HTML, el *Repositorio XML*, preservado con fines estadísticos y de control. Luego se agrega el repositorio de *Frases de roles* que son formaciones semánticas especificas para una taxonomía de roles.

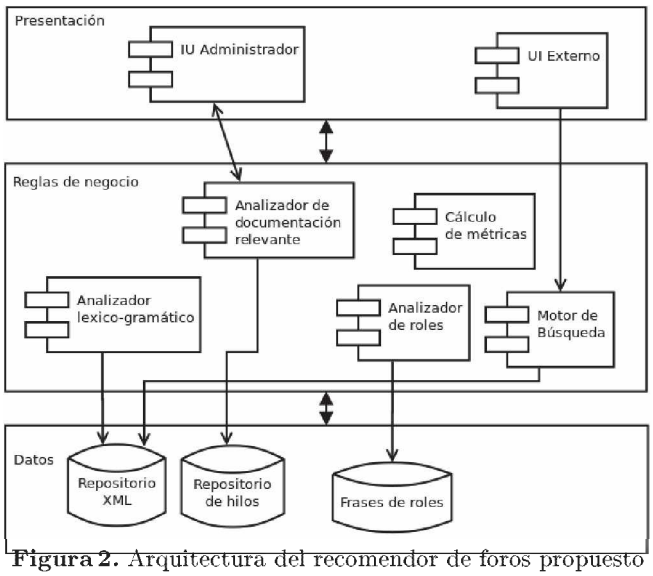

Figura 2. Arquitectura del recomendor de foros propuesto

La capa de Negocio de la aplicación incluye el componente *Motor de búsqueda,* encargado de la búsqueda y descarga de hilos de discusión disponibles en la web; mientras que el proceso de generación de documentos XML asociado se completa agregando etiquetas de algunos atributos puntuales y actualizando el repositorio de archivos XML. El componente *Analizador de documentación relevante* analiza la sintáxis de los fragmentos del hilo de discusión y los clasifica en base a la documentación Oracle Java [4]. Luego, el componente *Analizador léxico-gramático* mejora dicha clasificación mediante la incorporación de sinónimos utilizando WordNet de manera conjunta con la aplicación Stanford POS Tagger [5]. El componente *Motor de búsqueda* interactúa con la capa de presentación, solicitando al usuario que ingrese una cadena de búsqueda que represente su problema y, como resultado, devuelve una lista rankeada de soluciones candidatas, requiriendo los servicios del resto de los componentes que conforman las *Reglas de negocio.* El componente que se agrega en esta capa es el *Analizador de roles,* el cual será explicado a continuación en detalle.

La capa de Presentación incluye las interfaces de *Usuario Externo*, que provee una interfaz Web para que el recomendador de foros sea utilizado por usuarios interesados en encontrar soluciones a un problema técnico particular, y la de *Usuario Administrador,* para usuarios que realicen tareas de mantenimiento, carga y actualización de la base de datos, así como definición de nuevas métricas.

3. Analizador de Roles

El rol de un usuario participante está estrechamente relacionado con el conocimiento y expertitud del mismo; por consiguiente el componente *Analizador*

de roles tiene como función determinar los roles de los participantes de un hilo de discusión a partir de características y comentarios que contienen sus posts dentro de ese hilo. Para ello, partiendo de los hilos recuperados, se buscan frases o construcciones semánticas mantenidas en un repositorio de *Frases de roles*, identificándolas y clasificándolas, determinando así un rol para cada uno de los participantes de un hilo.

Se pueden diferenciar 5 roles distintos, los cuales se identifican como *Líder*, *Jefe*, *Moderador*, *Novato* y *Conflictivo*, cada uno de ellos puede ser identificado con un tipo de frases determinadas. Por ejemplo, los participantes con rol *Líder* son los usuarios con mayor conocimiento en el tema de discusión, emiten respuestas a las consultas, ayudan activamente y son reconocidos por sus aportes. Generalmente, luego de dar sus sugerencias, buscan verificar que sus respuestas hayan sido útiles. Se reconocen por comenzar sus posts con frases del tipo *"... lo primero que..."*,*"... prueba con este código..."*, *"... debes usar..."*, entre otras; y terminan sus post con frases del tipo *"... espero que te sirva..."*, *"... ya nos cuentas..."*.

Las personas con categoría *Jefe* tienen bastante conocimiento en el tema de discusión pero sólo dan directivas de solución. En general dan pautas, emiten links, sugieren búsquedas pero no resuelven los problemas. Los *Moderadores* conocen los temas y construyen lazos de comunicación entre usuarios, también supervisan; los usuarios *Novatos* tienen poco conocimiento sobre el tema de discusión; los usuarios *Conflictivos*, a diferencia de los otros roles, tienen comportamientos negativos en la conversación, emitiendo mensajes provocadores e irrelevantes en los temas que se discuten. Los usuarios *Novatos*, a su vez, tienen dos subroles bien determinados, el *Novato Implicado*, que si bien tiene poco conocimiento, a veces ayuda expresando la experiencia que tuvo él con un problema determinado y el *Novato Incipiente* tiene muy poco conocimiento y además no aporta ni contesta preguntas de otros usuarios. Los usuarios *Conflictivos* se categorizan a su vez en *Publicitario* (emite propagandas o información sin sentido), *Impostor* (amenaza o intenta engaños informáticos) y *Agresivo*, que utiliza lenguaje agresivo, grita (escribe en mayúsculas) o utiliza lenguaje indebido.

La Figura 3 esquematiza el proceso propuesto para la identificación de roles junto a la propuesta de evaluación de los resultados obtenidos. El proceso de *detección de frases de roles*, analiza el texto de los hilos e identifica las oraciones a partir del repositorio de *Frases por roles*, almacenado como parte de la taxonomía de roles. Mediante el proceso de *identificación de roles* se determina el rol final de cada usuario participante, los cuales están asociados a un nivel de conocimiento.

Por otro lado, para validar la metodología, para cada hilo se realiza un proceso de *clasificación manual* de los usuarios participantes, en base a una encuesta de opinión a un grupo de evaluadores humanos.

El objetivo del proceso de *comparación de resultados* es contrastar los resultados obtenidos mediante ambos procedimientos y proceder a su análisis en la etapa de *análisis de resultados*.

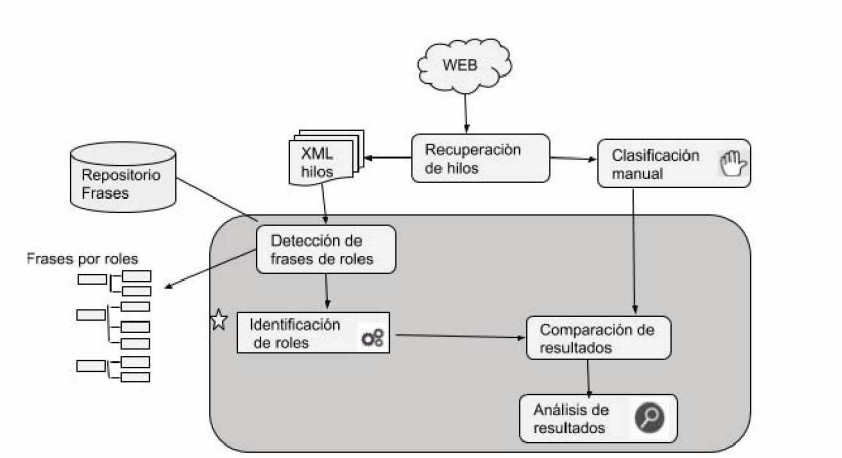

Figura 3. Esquema del proceso de identificación de roles en hilos de discusión

Detección de roles

Para poder detectar el rol que cumple un usuario dentro de un hilo de discusión se analiza el texto escrito por ellos en cada post utilizando como soporte el repositorio de frases definido anteriormente. Para ello, primero se definen las diferentes componentes que constituyen cada rol, como se puede apreciar en la Tabla 1. En dicha Tabla se observa que cada rol tiene un *id* asociado, un nivel de expertitud y un valor único (pR_i)que se encuentra dentro del rango de valores determinado para ese rol (*Rango según Rol* R_i.). Por ejemplo, en la primera fila, el *Rol* Líder – que corresponde al rol de mayor conocimiento –, tiene el nivel 1 de *Expertitud*, y como valor pR_i : 99 qué, en este caso, es el mayor del rango [90-99]. Esto se cumple en todos los casos, salvo para los subroles del rol *Conflictivo*, en los cuales se elige el menor valor del rango (puntajes 40, 30 y 20).

Cuando se analiza el texto del post de un usuario, se buscan frases que mapeen con las frases del repositorio. Luego de detectar dichas frases se puede saber a que rol pertenece un usuario en ese post, entonces se le asigna un *id rol* para el usuario en ese post y un valor pR_i, y así se hace con cada post perteneciente a ese usuario específico. Luego dicho valor pR_i es utilizado para poder detectar el rol global que cumple el usuario en el hilo de discusión.

La estrategia para detectar los roles de los diferentes participantes se basa en una fórmula que utiliza los valores obtenidos en cada post del participante, generando así un valor final por usuario. Este valor final se va a encontrar entre alguno de los rangos de roles establecidos anteriormente (Tabla 1) y va a determinar el rol que cumple un usuario dentro de un hilo de discusión. El valor mencionado se calcula a partir de la siguiente fórmula:

$$p(u) = \frac{\sum_{i=1}^{n}(c_i * pR_i)}{\sum_{i=1}^{n} c_i} \quad (1)$$

donde u es el Usuario que participa en un hilo de discusión de un foro; $p(u)$ es el puntaje de un usuario u; i es un escalar identificador del rol (entre 1 y 8);

Tabla 1. Asignación de puntos para roles de usuarios

Rol		Id rol R_i	Expertitud	Valor asignado (pR_i)	Rango según Rol R_i
Líder		R_1	1	pR_1=99	90–99
Jefe		R_2	2	pR_2=89	80-89
Moderador		R_3	3	pR_3=79	70–79
Novato	Implicado	R_4	4.1	pR_4=69	60–69
	Incipiente	R_5	4.2	pR_5=59	50–59
Conflictivo	Publicitario	R_6	5.1	pR_6=40	40–49
	Impostor	R_7	5.2	pR_7=30	30–39
	Agresivo	R_8	5.3	pR_8=20	20–29

c_i es la cantidad de veces que el usuario participa en el hilo de discusión con rol R_i; pR_i es el puntaje asignado al usuario con rol R_i, y n es la cantidad de roles/subroles posibles.

De esta manera, el puntaje establecido para el usuario u se obtiene a partir de considerar la presencia o no de cada uno de los 8 roles/subroles posibles producto el valor asignado a dicho rol (pR_i). Por último, el valor obtenido estará influenciado por la sumatoria todos los roles presentes. Para determinar el *valor global* se utiliza la columna *Rango según rol* de la Tabla 1 de acuerdo al valor obtenido $P(u)$.

4. Validación

Para estudiar el comportamiento de la propuesta en un contexto real, se llevó a cabo un caso de estudio [9]. Para ello se realizó un conjunto de encuestas basadas en hilos reales de 5 foros diferentes. La elección de los hilos se realizó considerando aquellos que tuvieran más de 8 posts, en los cuales intervengan más de 3 usuarios participantes en la discusión y cuya pregunta disparadora inicial sea una duda técnica. Para cada una de las encuestas se transcribió el hilo de discusión completo elegido, asignando al azar un conjunto de hasta 5 encuestas a cada uno de los evaluadores. Para cada encuesta un evaluador debía puntuar el grado de conocimiento de cada usuario en el hilo de conversación, siguiendo una escala numérica del 1 al 5, donde 1 equivale a mucho conocimiento; 2, a bastante; 3, a impreciso; 4 a poco conocimiento y 5 a no tener conocimiento en el tema. Los evaluadores fueron un grupo de 25 estudiantes de Informática (UNComa), 11 del primer año de la carrera de Licenciatura en Ciencias de la Computación y el resto del primer año de la Tecnicatura en Desarrollo Web, el detalle de la composición de las encuestas puede observarse en la Tabla 2 donde *Id* es el identificador de la encuesta, *Referencia URL hilo* es la dirección de la misma, *Cant.Eval.* indica la cantidad de evaluadores que completaron cada encuesta, *Cant.posts* corresponde con la cantidad de posts del hilo de discusión y la *Cant.Usu* es la cantidad de usuarios que participan del hilo de discusión.

Tabla 2. Resumen de encuestas

ID	Referencia URL hilo	Cant. Eval.	Cant. posts	Cant. Usu.
1	http://www.javahispano.org/java-se/post/1612752	14	9	4
2	http://www.espaciolinux.com/foros/programacion/ayuda-con-programa-proceso-hijo-ficheros-script-t34804.html	16	9	5
3	https://www.lawebdelprogramador.com/foros/Pascal-Turbo-Pascal/251768-contar-palabras-de-un-fichero.html	12	8	4
4	http://www.espaciolinux.com/foros/programacion/problema-con-shell-script-t32515.html	14	10	4
5	http://www.javamexico.org/foros/conceptos/el_punto_que_sicnifica	12	10	8

Para comparar los resultados obtenidos entre la heurística propuesta en la Sección 3 y el resultado de las encuestas, se cotejó el valor obtenido automáticamente , para cada usuario contra la media de los valores otorgados por los evaluadores a ese usuario, en el hilo correspondiente. El resultado puede visualizarse en la Figura 4, donde las barras representan el valor obtenido por la heurística y la línea continua representa la media otorgada por los evaluadores. Por ejemplo, el valor de la media otorgado al participante u_1h_1 (el cual corresponde al usuario 1 del hilo 1), concuerda entre el valor de la media $valormediaEvaluadores$ con R_4 (rol Novato), dado por la heurística. La misma concordancia se da para los usuarios participantes u_2h_1 , con el valor R_1 (rol Líder) , al usuario u_3h_2 con el valor R_3 (rol Moderdor) , usuario u_4h_2 con el valor R_2 (rol Jefe), usuario u_2h_3, con el valor R_2 (rol Jefe) y al usuario u_2h_5 con el valor R_3 (rol Moderador).

Al analizar en detalle los valores que difieren la media de los evaluadores y la heurística, se observa que la diferencia es de sólo 1 punto y que en pocos casos (6), se da una diferencia de 2 puntos, los cuales están detallados en la Tabla 3, donde se puede visualizar, para cada uno de esos usuarios, la conformación del $P(u)$, en la columna detalle, el valor de la media de los evaluadores y el de la heurística. Analizando específicamente los usuarios en los que hubo mayor diferencia entre la heurística y los evaluadores, se detectó que en el 83.4 % de los casos la media de los evaluadores asignó el valor 3, correspondiente a la opción de conocimiento *impreciso*.

Finalmente, al comparar los resultados obtenidos por los evaluadores con los hallados por la heurística en general se encuentra concordancia entre ambos resultados, en particular en los casos en que el grado de conocimiento estaba mejor definido. Otra detalle a destacar es que en muchos casos existe una

dificultad para determinar el nivel de conocimiento de los participantes cuando aparece un usuario participante con rol *Conflictivo*.

Tabla 3. Heurística vs media de evaluadores

Usuario	p(u)	Heuristica	Ruf	Detalle	Media evaluador
u_2h_2	49	5	R_6	$C_3 = 1$; $C_8 = 1$; $pR_3 = 79$; $pR_8 = 20$	3
u_4h_3	20	5	R_8	$C_8 = 1$; $pR_8 = 20$	3
u_4h_5	40	5	R_6	$C_6 = 1$; $pR_6 = 40$	3
u_5h_5	40	5	R_6	$C_6 = 1$; $pR_6 = 40$	3
u_6h_5	40	5	R_6	$C_6 = 1$; $pR_6 = 40$	3
u_8h_5	69	4	R_4	$C_4 = 2$; $pR_4 = 69$	2

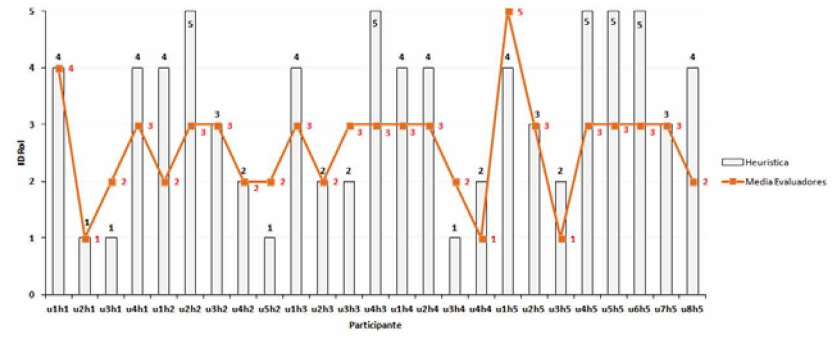

Figura 4. Heurística vs media de evaluadores

5. Trabajos relacionados

Otros trabajos han analizado los roles de los usuarios en hilos de discusión. Por ejemplo, Wever y otros [10] tienen como objetivo determinar la importancia de la asignación de roles de usuario para la construcción del conocimiento, para lo cual definen cinco roles (Started, Summariser, Moderator, Theoretician y Source Searcher), y utilizan el modelo de análisis de interacción Guanaardena [11], la diferencia es que lo utilizan en un entorno de aprendizaje colaborativo. Por otro lado, el trabajo de Lui & Baldwin [12] analiza los usuarios en base a cuatro habilidades básicas acordes al nivel de claridad de expresión en el post, de cuán positivo es el candidato, del esfuerzo de comunicación, del conocimiento, y la predicción de roles. Otra diferencia, es que estos autores utilizan redes de usuarios (grafos en donde los nodos son los participantes que emiten los posts y las aristas representan la clase de vinculación que hay entre los mismos), y redes de hilos (donde los nodos representan hilos de discusión y las aristas la similitud entre ellos) para su análisis. El trabajo, de Golder & Donath [13],

utiliza una taxonomía de roles y estrategias de interacción que describen el comportamiento que adoptan los usuarios en los foros de discusión para lograr obtener respuestas a sus consultas y, la diferencia con esta perspectiva, es que se enfocan en foros sociales de temas variados y gran volumen de información. Por su parte, Zhang y otros [14] definen su trabajo exclusivamente para la comunidad de Java, analizando el foro de discusión mediante una red social, clasificando cinco roles desde novatos hasta expertos Java. Su enfoque utiliza estadísticas costosas en términos de procesamiento, sumado a que se basa en la suposición de que si un usuario responde a otro indica un mayor conocimiento del que realizó la pregunta, lo que no siempre se cumple. Finalmente, la propuesta de Fisher y otros [15] utiliza la visualización y el análisis de redes sociales en los patrones de respuestas de cada autor, en grupos de noticias seleccionados, para encontrar diferentes tipos de participantes. Esto lo realizan por medio de redes egocéntricas calculadas sobre un usuario, usando el grado de entrada (a cuántas personas respondió) y grado de salida (cuánta gente le respondió) para identificar los roles dentro del grupo.

6. Conclusiones y trabajo futuro

En este trabajo se presenta una estrategia de clasificación de roles para participantes de hilos de discusión en foros técnicos, determinando para cada rol un grado de conocimiento y expertitud específico. Con el objetivo de evaluar la aplicabilidad de esta propuesta, se realizó un caso de estudio en base a cinco hilos de foros de discusión reales y se compararon los resultados obtenidos mediante la aplicación de la heurística con el análisis realizado por un grupo de evaluadores humanos. Al comparar los resultados , se detectó que en general los valores obtenidos mediante ambos métodos son similares. Específicamente, se encontró que en los casos que el grado de conocimiento de los usuarios está más definido (mucho, bastante, poco o ningún conocimiento) la diferencia fue mínima, y se presentaron algunas diferencias en los casos que los humanos clasificaron el nivel de conocimiento como impreciso. Dichos resultados apoyarían la propuesta por lo que se planea extender la cantidad de casos de estudio y replicar los experimentos realizados. A futuro, se planea incluir un proceso de retroalimentación para extender el conjunto de frases determinantes para cada rol y trabajar en la automatización de la misma.

Agradecimientos

Este trabajo está parcialmente soportado por el subproyecto "*Reuso de Conocimiento en Foros de Discusión, Parte II*", correspondiente al Programa de Investigación 04/F009 "*Desarrollo Orientado a Reuso, Parte II*", de la Universidad Nacional del Comahue (Neuquén, Argentina).

Referencias

1. G. Aranda, N. Martinez, P. Faraci, and A. Cechich, "Hacia un framework de evaluación de calidad de información en foros de discusión técnicos," in *ASSE 2013-*

Simposio Argentino de Ingeniería de Software, JAIIO 42Â°-Jornadas Argentinas de Informática, (Córdoba, Argentina), SADIO, 2013.

2. Aranda, Gabriela, Martínez Carod, Nadina, Roger, Sandra, Faraci, Pamela, and Cechich, Alejandra, "Una herramienta para el análisis de hilos de discusión técnicos," in *CACIC 2014, XX Congreso Argentino de Ciencias de la Computación*, (San Justo, Argentina), pp. 803 – 812, Oct. 2014.
3. N. Martínez Carod, P. Faraci, and G. N. Aranda, "Análisis de métricas de calidad en foros de discusión técnicos," in *XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).*, 2017.
4. V. Zoratto, N. M. Carod, F. Otermin, and G. N. Aranda, "Análisis de estrategias para clasificar contenidos en foros de discusión," in *XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).*, 2017.
5. F. Otermin, *Mejora de un recomendador de foros de discusión: Utilización de bases de datos léxicas para evaluación de sinónimos.* Licenciatura, Facultad de Informática, Universidad Nacional del Comahue, Neuquén, Argentina, Octubre 2018.
6. G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
7. "The stanford natural language processing group." Última modificación: 16-10-2018, Último acceso: 23-07-2019.
8. P. Faraci, *Evaluación de Calidad en Foros de Discusión Técnicos.* Licenciatura, Facultad de Informática, Universidad Nacional del Comahue, Neuquén, Argentina, Octubre 2017.
9. C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering.* Computer Science, Springer, 2012.
10. B. De Wever, H. Van Keer, T. Schellens, and M. Valcke, "Roles as a structuring tool in online discussion groups: The differential impact of different roles on social knowledge construction," *Computers in Human Behavior*, vol. 26, no. 4, pp. 516–523, 2010.
11. C. N. Gunawardena, C. A. Lowe, and T. Anderson, "Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing," *Journal of educational computing research*, vol. 17, no. 4, pp. 397–431, 1997.
12. M. Lui and T. Baldwin, "Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet," in *Proceedings of the Australasian Language Technology Association Workshop 2010*, pp. 49–57, 2010.
13. S. A. Golder and J. Donath, "Social roles in electronic communities," *Internet Research*, vol. 5, no. 1, pp. 19–22, 2004.
14. J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*, pp. 221–230, ACM, 2007.
15. D. Fisher, M. Smith, and H. T. Welser, "You are who you talk to: Detecting roles in usenet newsgroups," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, vol. 3, pp. 59b–59b, IEEE, 2006.